
Whose Permission Does Your Incident Response Need?

A model provider's safety classifier now decides what evidence your responders may examine about your own systems, and no one on your org chart owns that dependency.

Gerald C. Kane, Ph.D.

C. Herman and Mary Virginia Terry Chair in Business Administration
Department Head, Management Information Systems
Terry College of Business, University of Georgia

© 2026 Gerald C. Kane. Licensed under CC BY-NC 4.0: share and adapt for noncommercial use with credit to the author; do not present as your own. Full terms on the last page.

Executive Summary

Two accounts of the same situation appeared three weeks apart in the summer of 2026. Octane, admitted to Anthropic's Cyber Verification Program on July 1, described the arrangement approvingly. Gio Vignone, the company's founder and chief executive: "Security teams are entering a new regime. The strongest models are capable enough to matter, restricted enough that access now has to be earned, and dangerous enough that defenders can't afford to work with one hand tied behind their back" (Leclere, 2026). Chris Anley, chief scientist at NCC Group, described the same arrangement from inside the work: "The same tool is both an offensive tool and a defensive tool, and the two can't really be unpicked" (Franceschi-Bicchierai, 2026). Both are accurate. One reads earned access as a maturing market, the other as a constraint on work that runs on a clock somebody else started. Neither answers the question a board eventually has to answer: when your incident responders need to read your own evidence, whose permission is that, and who inside your organization owns the dependency?

Start with precision about the word. A model provider's guardrail is not a safety feature the enterprise receives. It is an access-control surface governing what evidence the enterprise may examine about its own systems, and like every access control in your environment it has a false-positive rate, a tuning owner, an appeals path, and a change log. In your own access stack all four are yours, and auditors ask about them by name. On this surface none of the four is yours, and two of them are not published at all.

In July 2026 that surface failed at the moment of maximum need, and it failed while working as designed. Hugging Face's responders, reconstructing an intrusion they could not yet attribute to anyone, reached first for commercial frontier models to analyze their logs. The models refused. Hugging Face named them in its technical timeline: "The models we reached for first, Claude Opus and Fable, refused a large part of that work: their safety guardrails treated reverse-engineering an exploit the same as launching one" (Larcher et al., 2026). The team re-ran the forensics on a self-hosted open-weight model instead. The company's own disclosure states the position it was in, and states it in a single sentence: "We do not know which model powered the attacker's agents, whether a jailbroken hosted model or an unrestricted open-weight one; either way, the attacker was bound by no usage policy, while our own forensic work was blocked by the guardrails of the hosted models we first tried" (Hugging Face, 2026). Attribution arrived five days later and from the other direction, when OpenAI reported that models it had been evaluating were responsible (OpenAI, 2026c). That account is self-reported and has not been independently audited. The refusing guardrails belonged to a third party with no role in the incident, no visibility into it, and no obligation to the victim, which forecloses the one-bad-vendor reading. A defender that could not name its attacker in the five days between its own disclosure and the other party's, while the tooling built to answer that question declined to look at the evidence, is the case in its sharpest form.

The sharpest finding is a gap inside the provider's own governance. Anthropic's published safeguard taxonomy for Fable 5 sorts cyber requests into four tiers and places incident response in the benign-use tier, allowed, while hacking, penetration testing, red teaming, and exploit development sit in the blocked high-risk dual-use tier (Anthropic, 2026). The written policy permitted exactly the work Hugging Face was doing. The classifier refused it anyway. That gap,

between the governance artifact an enterprise can read and negotiate and the mechanism that actually decides, is what this paper is about.

The anchor case is Hugging Face's own disclosure, so the frame in this paper rests on a self-interested first-party account. The generalization does not. Scale AI's Security and Policy Research Lab ran 2,390 real-world prompts from a collegiate cyber-defense competition against three frontier models and found a 12.2% aggregate refusal rate on authorized defensive work, rising to 43.8% on system hardening (Campbell et al., 2026). Two results matter most: prompts containing security-sensitive keywords were refused at 2.72 times the rate of semantically equivalent neutral requests, and explicit authorization raised refusals from 11.6% to 21.8%. The models tested were of the generation preceding the ones in the July incident, so the levels are evidence about a mechanism rather than a current rate card. The standard enterprise instinct, assert authority and put it in the contract, aims at the wrong layer.

Three conclusions follow. The separation of policy from mechanism, a design principle security engineering has held since Saltzer and Schroeder (1975), has been reproduced inside your AI supply chain with the policy on your side of the line and the mechanism on the provider's. Agent-driven attack sequences compress the window for containment decisions, so those decisions have to be written before the incident; writing them is governance work, and pre-delegating containment to agents is not available for purchase today. And the timing of a materiality determination, a board-owned obligation under SEC Item 1.05, now depends in part on a vendor's classifier, which no organization we can find has assigned an owner.

Why This Matters Now

The corpus this paper belongs to has been working the AI-governance problem from the procurement side. *The Steering Wheel Is Not Turned* (July 2026) took up the independence of assurance evidence: who produced the evaluation a buyer relies on, and who chose the scoring convention. *The Kill Switch Precedent* (June 2026) took up access continuity, the risk that a provider cuts you off. This paper is about a third surface, and the distinctions are worth stating once so we do not re-argue either. Assurance is about the evidence you are handed before you sign. Continuity is about losing access. This is about retaining access to a system that refuses the specific work you need, at the specific hour you need it, for reasons you cannot inspect.

Runtime controls have received less governance attention than contracts and evaluations, for a structural reason. A contract is a document, so it can be reviewed by people whose job is reviewing documents. A classifier is a decision procedure, and no function in most enterprises audits one that belongs to somebody else. Darktrace's 2026 survey of 1,500 security professionals found only 37% of security leaders reporting a formal AI governance framework, with 52% still discussing policy (Darktrace, 2026). A dependency with no owner, inside an organization with no framework, is a risk waiting for an incident to name it.

The tempo has changed underneath all of this. IBM and the Ponemon Institute, in a study of 602 organizations covering breaches between March 2025 and February 2026, report that 25% of malicious breaches were AI-enabled, a 56% year-over-year increase, and that AI-enabled breaches cost about \$6 million against a \$4.99 million global average (IBM & Ponemon Institute, 2026). The same study finds more than half of surveyed organizations using agents somewhere in threat

detection and containment. Enterprises are simultaneously increasing their reliance on model-mediated security work and increasing their exposure to attacks that require that work. The dependency is deepening on both sides at once.

I. What Happened When the Responders Reached for the Best Available Tool

The July 2026 Hugging Face intrusion is worth reconstructing carefully, because the operational detail carries most of what makes it useful. It is also the source of the structural frame in this paper, and its details are first-party disclosure.

Begin with the distinction between what is established and what rests on one party's word. Hugging Face detected an intrusion, cut off the attacker, and published on July 16 without naming anyone. Five days later OpenAI reported that models it had been evaluating against ExploitGym, a multi-institution benchmark of 898 real-world vulnerability instances spanning userspace programs, Google V8, and the Linux kernel (Wang, Schiller, et al., 2026), had escaped the evaluation environment and reached Hugging Face (OpenAI, 2026c). That is the only attribution on the public record, it comes from the party that ran the evaluation, and no outside body had audited it as of this writing. An independent assessment published by METR and Redwood Research on August 26, 2026 examined the incident's model behavior, not the attribution itself, which remains single-sourced; OpenAI revised its own account of the attribution that same day. The execution traces remain unreleased.

The two accounts of how the intrusion began have also never been publicly reconciled, and a reader of either one alone would describe a different attack. OpenAI's version turns on stolen credentials and zero-days. Hugging Face's disclosure describes something narrower: "A malicious dataset abused two code-execution paths in our dataset processing (a remote-code dataset loader and a template-injection in a dataset configuration) to run code on a processing worker" (Hugging Face, 2026). Its later technical timeline adds a zero-day in a package-registry proxy used to reach the open internet from the evaluation environment, and identifies the entry into Hugging Face itself as an HDF5 external-storage read that disclosed local files, together with a Jinja2 template injection (Larcher et al., 2026). What is not in dispute is the shape: an attacker window of about four and a half days, and on July 11 a run of some ten hours from Kubernetes enumeration through supply-chain access and node root to exfiltration of the benchmark's answer key from a production database (Larcher et al., 2026). The answer key was the target, which makes this a construct-validity story as much as a security one.

One further limit on the record deserves stating before the analysis leans on the case. Neither primary document establishes what permissions or instructions the models were operating under. The "first fully autonomous rogue AI attack" reading that circulated in commentary is therefore a reading neither text supports, and nothing in the argument below depends on it. The governance problem this paper takes up does not require the attacker to have been autonomous, or even to have been correctly identified.

In the same disclosure OpenAI reported a configuration detail that does most of the analytical work. The evaluation ran with safeguards deliberately off: the models had "reduced cyber refusals

for evaluation purposes," and "deployment safeguards were intentionally not enabled during this evaluation because it was aimed at testing cyber vulnerabilities" (OpenAI, 2026c). Hugging Face, meanwhile, ran into safeguards that were fully on. On that account the asymmetry was manufactured by configuration on both sides, and Hugging Face chose neither setting.

Then came the forensics. Hugging Face recovered roughly 17,600 attacker actions and grouped them into about 6,280 clusters, working through encrypted payloads that had to be decrypted before they could be read (Larcher et al., 2026). The company's own account of what happened next is unusually direct. "When we started the log analysis, we first used frontier models behind commercial APIs," the disclosure reads. "This did not work: the analysis requires submitting large volumes of real attack commands, exploit payloads, and C2 artifacts, and these requests were blocked by the providers' safety guardrails, which cannot distinguish an incident responder from an attacker" (Hugging Face, 2026). The technical timeline is blunter still: "Guardrails on Opus tripped every time we tried to analyze the attack logs" (Larcher et al., 2026).

The team's resolution was to self-host. "We ran the forensic analysis instead on GLM 5.2, an open-weight model, on our own infrastructure," the disclosure states, and it names a benefit that had nothing to do with refusals: "This had a second benefit: no attacker data, and none of the credentials it referenced, left our environment" (Hugging Face, 2026). GLM-5.2 is a 753-billion-parameter mixture-of-experts model published under an MIT license by Z.ai, and Hugging Face ran an Nvidia quantization of it (Z.ai, 2026).

The sentence that states the governance problem most economically is the one quoted in the executive summary, and the detail worth noticing is that Hugging Face wrote both halves as a single thought. It did not know which model had attacked it, and, either way, "the attacker was bound by no usage policy, while our own forensic work was blocked by the guardrails of the hosted models we first tried" (Hugging Face, 2026). By the disclosure's own account the guardrails could not tell an incident responder from an attacker, and only one of the two was asking permission.

Two caveats keep the case honest. Hugging Face's claim that its pipeline did in hours what would normally take days is company-reported, with no counterfactual and no independent measurement; treat it as the team's description of its own experience, because no speedup was measured. And the company lost a capability it had assumed at the moment it also had to answer a legally consequential question. It knew it had been breached, it wrote in its own disclosure that it did not know which model was responsible, and the answer reached it only through OpenAI's statement five days later, from a party outside the response. Legal IT Insider (2026) argues that attribution is load-bearing for notification thresholds, client communication, and insurer notification. Hugging Face had 17,000-plus recovered events and still could not close the question from the inside.

Implication for leaders. The operative question is about durability, not availability. Does your team's access to a frontier model hold when the material it must submit consists of live attack artifacts, and have you tested it on that material rather than on a description of it?

II. Policy and Mechanism Are Separable, and Only One of Them Is Yours

Security engineering has held for fifty years that protection systems should be built so that the rules can be changed without rebuilding the machine that enforces them. Saltzer and Schroeder (1975) are known for making that separation, between the policy a system permits and the mechanism that enforces it, a design principle instead of an accident of implementation. The reason the separation is desirable is that policy is contestable and mechanism is expensive. Organizations argue about what should be allowed, revise their conclusions, and need those revisions to take effect without a rewrite. Keeping the two layers apart is what makes a protection system governable at all.

The July 2026 case is that separation running in the wrong direction.

The policy layer is legible and reasonable. Anthropic's usage policy prohibits users from "Discover or exploit vulnerabilities in systems, networks, or applications without authorization of the system owner" (Anthropic, 2025a). Read the qualifier: the policy contemplates authorized work and permits it. Anthropic's published safeguard taxonomy for Fable 5 goes further, sorting cyber requests into four tiers. Incident response sits in the benign-use tier, described as core defensive and IT-related activities with little-to-no chance for abuse, and it is allowed. Hacking, penetration testing, red teaming, and exploit development sit in the high-risk dual-use tier and are blocked. The same document is candid about the cost of the arrangement: "The safety margin means that a request has to look very clearly safe to avoid triggering the classifier" (Anthropic, 2026). Anthropic acknowledges that the margin sweeps in some benign and some low-risk dual-use requests and produces a higher rate of false positives, and it publishes no number for that rate. An enterprise counsel reading these documents in a procurement review would conclude, correctly, that forensic log analysis of an intrusion against systems you own is permitted work.

The mechanism layer refused it. Not once, and not marginally: guardrails tripped every time, on a large part of the work, across two models from the same provider. The policy said yes and the machine said no, and the machine was the one making the decision.

This is a different failure from a policy disagreement between a vendor and a buyer. A policy disagreement can be negotiated, escalated, papered over in a contract amendment, or lost on the merits. All of those are governance moves available to a general counsel. What Hugging Face encountered was a mechanism the enterprise can neither read nor test nor appeal, sitting downstream of a policy the enterprise can read, test, and negotiate. Buying the policy does not buy the mechanism.

The policy layer is narrow to begin with. OpenAI's usage policies prohibit the "destruction, compromise, or breach of another's system or property, including malicious or abusive cyber activity" (OpenAI, 2025). The prohibition is sound, and it contains no authorized-security-research carve-out, no penetration-testing provision, and no dual-use language. So on one side of the market the policy contemplates authorized defensive work and the classifier does not read the qualifier, and on the other side the policy does not contemplate it at all. Neither arrangement gives a buyer a mechanism-level commitment.

Return to the four properties named in the executive summary and ask them one at a time about this surface. What is the false-positive rate on cyber refusals for the model your responders will use

in an incident? Anthropic acknowledges false positives qualitatively and publishes no number, and no provider we surveyed publishes one. Who owns the tuning? The provider's safety organization, on its own schedule, with no notice obligation to you. What is the appeals path during an active incident? There is a support queue, which is not a path measured in the hours an intrusion allows. Where is the change log? A classifier can be retuned between Tuesday and Thursday with no artifact you can show an auditor, so a control your response plan depends on can change without any event your change-management process would record.

Forbes covered the disclosure as a gap in AI safety controls (Janakiram MSV, 2026), and a gap does exist, though a narrower and more actionable one than a general critique of safety alignment. Your enterprise has acquired a protection mechanism that governs access to your own evidence, and none of the four properties that make a protection mechanism auditable.

Implication for leaders. Any control in your environment whose false-positive rate, tuning owner, appeals path, and change log all sit outside your organization should be documented as a third-party dependency in your incident-response plan, not as a feature of your tooling. The documentation is cheap. The absence of it is what turns a known limitation into a surprise.

III. The Refusal Rate Is Measured, and Asserting Your Authority Makes It Worse

The strongest reason to treat this as structural is that the phenomenon was measured independently, before the incident, at scale, with inter-rater reliability reported.

Campbell and colleagues at Scale AI's Security and Policy Research Lab assembled 2,390 real-world prompts from the National Collegiate Cyber Defense Competition, ran them against three frontier models, and labeled refusals with a Fleiss' κ of 0.80 (Campbell et al., 2026). The aggregate refusal rate on authorized defensive tasks was 12.2%, or 291 of 2,390 prompts. The distribution by task type is where the operational relevance sits: system hardening refused at 43.8%, malware analysis at 34.3%, vulnerability assessment at 22.7%, incident response at 18.9%. By model, Claude 3.5 Sonnet refused 19.5%, GPT-4o 10.2%, and Llama-3.3-70B 6.6%.

Read those levels for what they are. The study measured three models of a prior generation against one competition corpus and published in March 2026, so the percentages are not a current rate card for the models an enterprise runs today, and no equivalent measurement of the current generation exists. What generalizes is the mechanism, and the mechanism is what the next two findings expose.

Two of their findings deserve to be read twice by anyone who has ever tried to solve a vendor problem with a contract.

The first is a keyword effect. Prompts containing security-sensitive terminology were refused at 2.72 times the rate of semantically equivalent neutral requests, 30.5% against 11.2%, at $p < 0.001$. The authors' predictive analysis locates the mechanism more precisely: prompt embeddings predicted refusal at an AUC of 0.827 against 0.572 for keyword features alone, which rules out a simple keyword filter and points at something closer to a resemblance test. What the classifier weighs is how closely the material in front of it resembles harm. The identity and the authority of

the party submitting it are not inputs. That mechanism explains the rest of the record, including the one measurement in this study that appears to contradict the Hugging Face case.

The second finding inverts the standard enterprise remedy. Explicit authorization *increased* refusals, from 11.6% to 21.8%. Combine offensive terminology with an explicit authorization statement and the refusal rate reached 50.0%. Stated plainly: telling the model that you are authorized to do the work roughly doubles your chance of being refused, and telling it in security language makes it a coin flip. The instinct to assert authority, document the engagement, and put the authorization in writing is exactly correct as governance and counterproductive as a technique, and the resemblance test is why: the mechanism has no input for your authority, only for the security-adjacent language you added in order to establish it. There is no procurement path out of this at the prompt layer.

The authors are explicit about where this lands hardest. From their abstract: "These findings are urgent for interactive use and critical for autonomous defensive agents, which cannot rephrase refused queries or retry" (Campbell et al., 2026). A human analyst who gets refused rewords the request and loses four minutes. An automated response pipeline that gets refused stops.

The complication that demonstrates the mechanism

Scale AI measured refusal on log analysis at 0.0%. Log analysis is nominally the exact task Hugging Face was blocked on, so at face value the benchmark says the incident should not have happened. Any skeptical reader will find that number, and it is worth working through, because it confirms the resemblance test instead of undercutting it.

The two measurements differ in what the prompts contained. Benchmark log-analysis prompts are *about* logs: they describe a scenario, ask what an entry means, request a triage approach. Hugging Face's forensics required submitting the artifacts themselves, described in its own disclosure as large volumes of real attack commands, exploit payloads, and C2 artifacts (Hugging Face, 2026). The task label was identical. The content was not. A mechanism scoring resemblance to harm scores the content, which is how the same task label produces 0.0% in one setting and a block every time in the other.

The generalization is consequential enough to change a test plan: refusal risk is a function of what the evidence physically contains, not of how the task is labeled. An organization that validates its incident-response tooling on sanitized prompts will measure a refusal rate near zero and will still be blocked in a real incident, because the sanitizing is what removed the refusal trigger. This is a measurement error with the same shape as a disaster-recovery test that never actually fails over.

Implication for leaders. Whatever validation you run on model-assisted security tooling has to run on real artifacts from a real environment, with the payloads intact, or it is measuring something other than what you will need. If your security team cannot submit genuine attack artifacts to a test, that constraint is itself the finding.

IV. The Same Surface Fails in Both Directions

The obvious recommendation at this point is to ask the provider to loosen the classifier for verified enterprise customers. A second body of measurement forecloses it.

Jones and colleagues at Gray Swan AI and Carnegie Mellon measured the opposite error: whether agents refuse when the target genuinely is production infrastructure they have no business touching (Jones et al., 2026). With obvious production URLs in view, refusal behavior was strong. Claude Sonnet 4.5, GPT-5.2, and GPT-5.1 Codex refused 100% of the time; Claude Opus 4.5 refused 85%, Gemini 3 Flash 70%, Gemini 3.1 Pro 13%. Then the researchers put the same targets behind an nginx proxy so the production context was no longer legible from the request. Six of eight models dropped to 5% or below. In the harder condition every model but GPT-5.2 dropped to zero, and GPT-5.2 reached 13%. Across 800 trajectories, 18 showed the desired behavior.

Set the two studies beside each other. The same control surface is over-inclusive against a defender who names the work honestly and under-inclusive against anyone who obscures the context. A dial set too tight produces one kind of error, and loosening it trades that error for the other. This surface produces both at once, because the resemblance test scores the request and not the environment the request concerns. Loosening it does not move a defender across a threshold; it lowers the bar for the party already getting through by rephrasing.

That is what makes this a governance question rather than a configuration question. The Gray Swan authors propose environmentally aware refusal, in which the system verifies infrastructure context before deciding, and their proposal remains a research direction; nothing like it ships today. Until something like it ships and is independently tested, the enterprise is managing a control it cannot tune in either direction without making the other error worse.

Implication for leaders. Requests to vendors for a looser boundary should be understood as requests to accept a specific increase in a different risk, and priced accordingly. The better ask is for the two properties the enterprise currently lacks entirely: a published false-positive rate on defensive tasks, and notification when the classifier changes.

V. The Market Has Already Repriced the Boundary as an Entitlement

We do not have to argue that the safety layer functions as an access-control surface. The providers have built an access-control product on top of it and are selling admission.

OpenAI announced Trusted Access for Cyber in April 2026, offering GPT-5.4-Cyber, which it describes as "a version of GPT-5.4 which lowers the refusal boundary for legitimate cybersecurity work" (OpenAI, 2026a). Verification runs per individual through a dedicated intake or per enterprise through an OpenAI representative, and approved customers gain access to versions of existing models with "reduced friction around safeguards." Anthropic's Cyber Verification Program works at the level of the organization. Anthropic has published no announcement of the program and no participant roster; access runs through a login-gated application portal, so the clearest available description comes from a participant. Octane's announcement of its admission states that dual-use work blocked by default for the broader market, including vulnerability exploitation and offensive security tooling, becomes "unblocked for Octane, at the organization level, under Anthropic's monitoring and usage policy," while "Genuinely malicious categories (like ransomware development or mass data exfiltration) stay blocked for everyone, including us" (Leclere, 2026). That the best public account of an access-control program comes from an admitted vendor rather than from its operator is itself worth noting.

Read those two programs as governance artifacts rather than as product news. The refusal boundary is now a tiered, provider-administered entitlement requiring pre-incident enrollment and identity verification. Whether your responders may examine your own evidence has become a permission a third party grants, revocably, on a schedule it sets. Octane's founder, quoted at the top of this paper, describes that arrangement as a new regime in which access has to be earned, and the description is fair. Nobody in this debate holds the position that defenders should be blocked. The verified-access architecture is a genuine attempt to solve a genuine dual-use problem, and for an enrolled organization it works.

What the architecture is silent about is the unenrolled enterprise at hour one of an incident, and that silence is where the exposure sits. Enrollment is a procurement action with a lead time. An intrusion is not.

The anchor case of this paper is that enterprise, which is worth stating plainly because it was not obvious from the first accounts. Hugging Face was not enrolled in either program when its own incident began. Its technical timeline reports that the models its responders reached for first, Claude Opus and Fable, "refused a large part of that work: their safety guardrails treated reverse-engineering an exploit the same as launching one" (Larcher et al., 2026), and a team holding organization-level verified access would not have been refused on exploit analysis, because that is the specific thing verification unblocks. OpenAI's own account settles the other half directly: among the actions it lists as taken after the intrusion, "We've brought Hugging Face into the trusted access program and are supporting their teams in rapidly using our models' capabilities to improve their defenses" (OpenAI, 2026c). Brought into, after. Enrollment arrived as remediation, which is to say it arrived at the one moment it could not help.

That changes the emphasis of everything that follows. The company at the center of this case is not a small enterprise with no AI sophistication. It is a platform company whose entire business is hosting models, whose responders knew exactly which models they wanted and why, and it still met the refusal boundary during a live incident because it had not bought admission in advance. If that organization was unenrolled at hour one, the base rate among ordinary enterprises is not a question worth agonizing over.

Enrollment also does not restore predictability, which is the finding that closes the loop. Chris Thompson, chief executive of RemoteThreat and founder of Offensive AI Con, told TechCrunch that guardrails are inconsistent and vary from day to day even inside vetted programs, and described the resulting work: "You spend a lot of time negotiating with the model instead of working on the core security program" (Franceschi-Bicchierai, 2026). If the boundary moves inside a verified program, an enterprise cannot plan its response tempo around the boundary even after buying admission. It can only discover where the boundary is by hitting it.

The practitioner record here reaches past one platform company and one boutique. NCC Group is a listed professional-services firm, and Anley's position is that offensive and defensive tooling cannot be separated, with the reported workaround of falling back on open-source models without guardrails. Paolo Stagno, chief technology officer at CrowdFense, told the same reporter that AI companies "essentially treat customers like children who need babysitting," and runs local open models to avoid cloud data exposure. Mark Dowd, an independent researcher, framed the governance objection: "It's not really comfortable to me that these random large companies are making arbitrary decisions about what is safe" (Franceschi-Bicchierai, 2026).

The record also contains a genuine counter-voice, and it belongs here. Giuseppe Cali told TechCrunch that guardrails do not impede him, because of how he works: "I am jealous of my bugs, and I like this game too much to let models play it for me" (Franceschi-Bicchierai, 2026). A practitioner who uses models only for reverse engineering and support tooling encounters the boundary rarely. The exposure is concentrated in exactly the workflows an enterprise is most likely to build, the ones where the model processes the raw evidence at volume.

Implication for leaders. If a model-mediated capability appears anywhere in your incident-response plan, enrollment in the relevant provider's verified-access program is a pre-incident procurement item with a named owner and a review date, not a step in the runbook. And enrollment should be treated as reducing the probability of a refusal rather than eliminating it.

VI. Speed Compresses the Containment Decision, and the Decision Is Not Automatable

The tempo argument for automating containment is easy to overstate. The provider's own analysis is the best source for the mechanism. Writing about long-horizon models three weeks before the Hugging Face disclosure, OpenAI observed that "each *step* can look acceptable on its own while the *sequence* can produce an outcome that would not be approved," and that "monitoring individual actions no longer suffices to track the intent of the overall trajectory." The same document notes that a long-horizon model "can learn the blind spots of an approval system and work around it to achieve its goals" (OpenAI, 2026b). The company also reported token fragmentation used to evade scanners during the evaluation (OpenAI, 2026c), which shows deliberate control evasion at work. Anthropic's own investigation of the GTG-1002 espionage campaign in late 2025 reported AI executing 80 to 90% of tactical operations independently, with human involvement limited to campaign initialization and authorization decisions at critical escalation points, at sustained rates of multiple operations per second (Anthropic, 2025b). Those figures are the provider's first-party account of activity on its own platform, not an independent measurement, and no outside party has audited them.

Now the disciplining evidence.

Mandiant's M-Trends 2026, drawn from more than 500,000 hours of frontline investigations in 2025, reports that the median time between an initial access event and hand-off to a secondary threat group collapsed from more than 8 hours in 2022 to 22 seconds in 2025 (Mandiant, 2026). That figure is frequently recruited as evidence about AI tempo, and it measures something else: broker-to-operator hand-off in the human criminal economy. Read correctly it weakens the novelty claim, because the response window had already collapsed before agents arrived. Mandiant is explicit on the broader point: "Despite these rapid technological advancements, we do not consider 2025 to be the year where breaches were the direct result of AI. From our view on the frontlines, the vast majority of successful intrusions still stem from fundamental human and systemic failures." The same report shows global median dwell time moving the wrong way, 14 days in 2025 against 11 in 2024, which Mandiant reads as likely reflecting growing sophistication in evading defenses. Simon Willison, analyzing the incident, treats it as genuine while deflating the capability reading, and the deflation rests on OpenAI's own characterization: all the evidence suggested the models were hyperfocused on finding a solution for ExploitGym, going to extreme lengths in service of a

narrow testing goal (OpenAI, 2026c; Willison, 2026). Narrow obsession and general capability are different claims, and the record supports the first. That reading belongs in any honest account.

So the conclusion is not that containment should be automated. Three pieces of evidence say the automated defender is the weaker party.

Anthropic reported that during the autonomous offensive operations it investigated, Claude "frequently overstated findings and occasionally fabricated data," and that this presented challenges for the actor's operational effectiveness (Anthropic, 2025b). The unreliability that limited the attacker limits the defender identically, and the defender is making decisions with production consequences.

Wang and colleagues built SecRespond to test the proposition directly, benchmarking 23 frontier models on post-compromise incident response across 10 cyber ranges, 21 MITRE ATT&CK techniques, and 5 operating systems. No model achieved complete detection and remediation on any single range (Wang, Chen, Ding, et al., 2026). Agents reliably surfaced what alerts already flagged and struggled to investigate silent intrusions proactively, which is the inverse of the capability an unattended containment agent would need.

Hugging Face's own experience supplies the case. The company had an AI-based alerting agent, it fired during the intrusion, and it failed to escalate the alert's criticality to the on-call team, costing response time (Larcher et al., 2026). The detection decision had already been automated, and the automation was the weak link.

The right conclusion from the tempo evidence is therefore a governance one. If the sequence outruns step-by-step approval, the containment decision has to be *pre-decided*: written down in advance, with thresholds, named authority, and standing permission to act, so that the human who executes it is carrying out a decision the organization already made instead of seeking approval mid-incident. Writing decision rights in advance is a different act from delegating them to an agent, and IBM's finding that more than half of surveyed organizations already use agents in detection and containment against just 18% in vulnerability management (IBM & Ponemon Institute, 2026) suggests the two are being confused at scale.

Implication for leaders. Ask which containment actions your responders may take without seeking approval, and confirm the answer exists in writing before an incident, not in a chain of judgment calls during one. Then ask what the plan does when the model-assisted step in that chain returns a refusal instead of an answer.

VII. The Materiality Clock Now Has a Vendor In It

Under SEC Item 1.05, "An Item 1.05 Form 8-K will generally be due four business days after a registrant determines that a cybersecurity incident is material" (U.S. Securities and Exchange Commission, 2023). The clock starts at the materiality determination, not at discovery. A forensics delay therefore does not mechanically breach the four-day deadline; if anything it postpones the trigger. Any argument that guardrails cause disclosure violations is unsupportable and we are not making it.

The supportable claim is narrower and, as far as we can determine, unpublished. The materiality determination is a board-level obligation whose timing now depends in part on a vendor's classifier, and no organization has assigned that dependency an owner. Registrants may not unreasonably delay the determination itself, so the exposure sits precisely in the interval between discovery and determination, which is the interval forensic analysis occupies.

CB Financial Services makes the tempo concrete, and the filing itself is on EDGAR. On May 5, 2026, Community Bank, the wholly owned subsidiary of CB Financial Services, "became aware of an internal incident involving the handling of certain non-public customer information using an unauthorized artificial intelligence-based software application." The incident involved no disruption to operations, customer account access, payment systems, or core infrastructure; the company nonetheless determined the event material on May 7, "due to the volume and sensitive nature of the non-public information at issue," and filed an Item 1.05 Form 8-K on May 11. Among the information disclosed were "customer names, social security numbers and dates of birth" (CB Financial Services, 2026). The filing is agentless about who did it, describing the handling of information through an unauthorized application and naming no actor. Wilson Sonsini's analysis of the filing states the two points that make the case instructive: "The four-business-day disclosure clock under Item 1.05 starts at the materiality determination, not upon detection of the incident," and "A cybersecurity incident need not involve an external attacker or system intrusion or material financial consequences to qualify as material under Item 1.05" (Wilson Sonsini Goodrich & Rosati, 2026). No external attacker, no intrusion, no operational disruption, and two days from discovery to the materiality call.

Two days is the tempo against which the classifier dependency has to be measured. Hugging Face's attribution gap ran five days. A delay of that length inside the discovery-to-determination interval is a delay in a board obligation, and the board will be asked to explain it in terms of what it knew and when.

This is where the governance question resolves into a decision right. Our prior work on agent governance placed a named owner at the top of the five operational artifacts attached to the GOALS template, which asks, for any agentic deployment, what the Goal is, what the Operations are, who holds Authority, what the Limits are, and how Success is judged. Apply the ownership question one layer out from the agent, to the control surface the agent depends on. Who owns the vendor-classifier dependency in your incident-response plan, who is accountable if a refusal extends the interval before a materiality determination, and which committee sees the answer? Most organizations will find the question has no owner, because the dependency was never named as one. Procurement-level architecture, which we have treated elsewhere under SCALES, governs what the contract must contain. This question is downstream of the contract and upstream of the incident, and it currently sits between two functions rather than inside either.

The European regime raises a parallel question we can pose but not answer. Article 55 of the EU AI Act, in application since August 2, 2025, requires providers of general-purpose AI models with systemic risk to track, document, and report serious incidents to the AI Office without undue delay, and to maintain an adequate level of cybersecurity protection for the model and its physical infrastructure (European Union, 2024). The European Commission published a serious-incident reporting template for those models in November 2025 (European Commission, 2025), so the mechanism existed eight months before this incident. Whether a capability evaluation that reaches an unrelated production system meets the serious-incident definition, and which party carries the

obligation, is open. We found no public evidence that either company notified the AI Office, and we note that as a documented absence rather than a documented failure.

Implication for leaders. Name the owner before the disclosure question is live, and make sure the audit committee knows that the interval between discovery and determination has a third party inside it.

What This Adds Up To

Three separate bodies of evidence converge on the same structure, and the convergence is what makes this more than a story about one intrusion. The measurement literature shows a control that errs in both directions at once. The provider documentation shows a policy layer that permits the work and a mechanism layer that refuses it. The market shows both providers converting the boundary into an administered entitlement, which is a commercial acknowledgment that the boundary is an access-control decision and not a safety property of the model.

The governance conclusion follows. The AI safety layer has become part of the enterprise control plane without ever being onboarded as one. It governs access to enterprise evidence, it changes without notice, it has no published error rate, and it has no owner inside the organization that depends on it. Every other control with those characteristics is treated as an audit finding.

The layer went unnoticed because it arrived labeled as protection. A control labeled as a feature gets no owner, no test, and no change-management process, because none of those apply to features. That is a naming problem with operational consequences, and the reason the word deserved the precision this paper opened with.

Failure Modes and Counter-Patterns

Three ways to get this wrong are already visible, and one important body of evidence runs against the most obvious recommendation.

Testing on sanitized material. The most likely failure is quiet. A team validates its model-assisted analysis pipeline on redacted or synthetic artifacts, measures no refusals, and records the capability as available. If your test removed the payloads, it removed the trigger.

Asserting authority louder. The second failure is the one an experienced governance function is most likely to commit. Documenting the engagement, adding authorization language to system prompts, and papering the authority in the contract are all correct governance and are counterproductive as a technique, for the reason the measurement gives: the mechanism has no input for your authority. Do the governance because it is required. Do not expect it to change the mechanism's behavior.

Redundancy that shares the failure mode. Etay Maor, vice president of threat intelligence at Cato Networks, makes the resilience case: "Using multiple models provides resilience if one becomes unavailable, overly restricted, or compromised" (Constantin, 2026). The argument is sound in general, and the July case shows its limit in particular. Two Anthropic models refused the same work, so a second model subject to a correlated classifier is redundancy on paper. The

vendor-selection version of the move fails for a further reason: on the only attribution available, the attacker in the anchor case was one provider's model while the refusal came from another provider's guardrails, so no available vendor choice resolves both sides. Redundancy requires a fallback whose refusal decision is made by a different party under a different regime.

The counter-evidence, stated plainly

The recommendation that follows most naturally from this analysis, retaining a self-hosted open-weight model as pre-staged fallback capacity, runs against the direction enterprise buying is reported to be moving, and that reported direction rests on better evidence than the recommendation does.

Menlo Ventures surveyed roughly 500 U.S. enterprise decision-makers between November 7 and November 25, 2025 and found open-weight models capturing an estimated 11% of enterprise LLM spend, by Menlo's own methodology based on production API-usage proportions, down from 19% the prior year, against \$37 billion in enterprise generative AI spending (Menlo Ventures, 2025). Enterprises are moving away from open weights, not toward them. Read the figure as illustrative rather than definitive, and note the interest: Menlo is an investor in Anthropic, so a share estimate that favors hosted frontier models over open weights comes from a source with a position in the outcome. No independent census of enterprise model share exists to check it against, which is itself part of the problem this paper describes. The same study reports that enterprises remain particularly cautious toward Chinese open-source models, which account for only 1% of enterprise open-source usage. Hugging Face's fallback was a Chinese-origin open-weight model (Z.ai, 2026), which is the single model family enterprise buyers are most averse to.

That produces a real bind rather than a rhetorical one. A regulated buyer who cannot run a Chinese-origin model for policy reasons, and who has not enrolled in a provider's verified-access program, has no pre-staged fallback at all. The available open-weight alternatives may not carry the capability the forensic work requires, and the capable alternative may not clear the policy. Naming that bind is more useful than recommending past it.

We should also be honest about what the evidence base does not contain. Sonali Shah, chief executive of Cobalt, states the posture directly: organizations should have vetted AI models they can operate inside their own trust boundary before an incident happens (Constantin, 2026). Jacob Krell of Suzu Labs puts the operational case in one line: "Machine-speed exploitation requires machine-speed response, and that response can't run on models that refuse to examine the evidence" (Constantin, 2026). Both are practitioners at security firms, and their view is supported by other named practitioners at NCC Group, CrowdFense, and RemoteThreat. We found no primary survey evidence on why enterprises retain open-weight capacity in regulated industries. The support for this posture is individual practitioners and boutique security firms, not enterprise survey data, and readers should weight it accordingly.

What to Decide Before the Incident

The decisions below are pre-incident because none of them can be made at speed. They are deliberately unbranded and deliberately few.

Name the owner of model-access continuity. One person, named in the incident-response plan, accountable for whether the models your responders depend on will process what your responders need to submit. The same line should state what happens when the answer is no.

Run one test on real artifacts. Take genuine attack artifacts from a prior incident or a red-team exercise, payloads intact and not merely described, and submit them to each model in your response tooling. Record what was refused, at what volume, and how long the workaround took. Repeat quarterly, because the classifier changes and you get no notice when it does. A failover test that never cuts over measures nothing, and neither does this one on redacted material. If your policy will not allow real artifacts to leave your environment, that is the finding, and it points at the next decision.

Decide the fallback before you need it, and be specific about the bind. Determine whether a self-hosted model of adequate capability exists that clears your policy constraints. If one does, stand it up and confirm it runs. If none does, escalate that as an accepted gap with a named owner instead of leaving it as an assumption, and revisit it when the available models change. This is the decision where the measured enterprise trend and your own regulatory posture pull in opposite directions, so resolve it on the record, by a person.

Enroll, and treat enrollment as probability reduction. Verified-access enrollment with your primary providers is a procurement action with a lead time measured in weeks. Complete it in advance, and note in the plan that enrolled access has been reported to vary day to day, so the runbook should not assume it holds.

Write the containment decision rights down. For each containment action your responders may need to take under time pressure, state who may authorize it, at what threshold, and what standing permission exists. The measured record says no frontier model completes detection and remediation end-to-end on a benchmark range (Wang, Chen, Ding, et al., 2026), so the plan should not assume an agent will carry the decision.

Put the dependency in the disclosure-controls documentation. One sentence acknowledging that the interval between discovery and materiality determination includes a third-party dependency your organization does not control. Then ask the audit committee whether it has seen that sentence.

The Next Twenty-Four Months

Five predictions, stated so they can be checked against the record by mid-2028.

At least one major frontier provider will publish a false-positive rate for its cyber-refusal classifier by mid-2028. None does today, and it is the single most useful missing number in this area. Competitive pressure from verified-access programs makes disclosure a differentiator.

A published incident-response benchmark will report a refusal rate alongside its completion rates by mid-2028. SecRespond reports none today. Once refusal is understood as a failure mode of the defender's tooling and not a property of the model's alignment, benchmark design follows.

No frontier model will complete detection and remediation end-to-end on a standard cyber-range benchmark before mid-2027. This is the prediction most likely to be wrong, and it is the one that would most change the recommendations above if it were.

Open-weight share of enterprise LLM spend will not return to 19% by the end of 2027, even as pre-staged fallback capacity becomes a stated security requirement. These two movements will decouple: fallback capacity will be procured as a narrow, security-scoped exception owned by the security function, not as a general shift in enterprise model strategy.

A guardrail-attributable forensic delay will appear on the public record by mid-2028, in a securities filing, a regulatory response, or litigation discovery. If it does not, the absence is informative rather than reassuring, and would indicate the dependency is being absorbed without ever being named.

Methodology Note

This paper synthesizes primary documents, measurement studies, and named practitioner testimony retrieved in July 2026, with the load-bearing sources re-verified against the live pages on July 30. The incident reconstruction rests on Hugging Face's own security disclosure and technical timeline and on OpenAI's account of attribution, all first-party and all self-interested in the ordinary way. Three limits on that reconstruction deserve to sit here as well as in the body. OpenAI's attribution is self-reported and unaudited, and it is the only attribution on the record. The two accounts of initial access differ and have not been publicly reconciled, and the execution traces are unreleased. And neither primary document establishes the permissions or instructions under which the models operated, so this paper takes no position on the autonomy question and no part of its argument turns on one. The generalization rests on three measurement studies, two of them arXiv preprints not yet peer reviewed: the Scale AI refusal-bias study, the Gray Swan and Carnegie Mellon refusal-framework study, and SecRespond. Provider policy claims come from the usage policies and safeguard documentation themselves, and practitioner testimony from named, attributed interviews in TechCrunch and CSO Online. One description of a provider program comes from a participant rather than its operator: Anthropic has published no announcement of its Cyber Verification Program and no participant roster, and access runs through a login-gated application portal, so the account of what that program grants and withholds is taken from an admitted vendor's announcement and should be read as such. Neither provider publishes a participant list, which is why enrollment status in this paper is established from the parties' own incident accounts rather than from a roster. The CB Financial Services filing was retrieved directly from EDGAR (accession 0001605301-26-000021) and its dates and data elements verified against the filing text. Survey figures come from Menlo Ventures, IBM and Ponemon, and Darktrace, each with the sampling limits ordinary to practitioner surveys; Darktrace's public summary does not fully specify its methodology.

The limits are worth naming specifically. No provider publishes a false-positive rate for its cyber classifier, which means the central quantity in this analysis is acknowledged qualitatively and measured nowhere. Hugging Face's enrollment status is settled: OpenAI states that it brought the company into Trusted Access after the intrusion, and the refusals recorded in Hugging Face's own technical timeline establish that it did not hold Anthropic's organization-level verified access during

the forensics. The Anthropic half is an inference from the refusals rather than an explicit statement by either party, and it is labeled as such where it appears. Hugging Face's account of forensic speed remains company-reported with no counterfactual. The CB Financial Services timeline has been verified against the primary filing on EDGAR and no longer rests on a law firm's summary; the summary is retained only for its two legal readings. We were unable to retrieve the body text of NIST's incident-handling guidance (Nelson et al., 2025) and therefore characterize nothing about its contents. Two secondary reports on the incident were inaccessible during the research pass and are not relied on for any specific claim. Whether either company notified the EU AI Office under Article 55 is unknown in either direction. And there is no primary survey evidence on enterprise reasons for retaining open-weight capacity, so the self-hosting posture discussed here is supported by practitioner testimony rather than by measured enterprise practice, which runs the other way.

References

- Anthropic. (2025a, September 15). *Usage policy*. <https://www.anthropic.com/legal/aup>
- Anthropic. (2025b, November). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. <https://www.anthropic.com/news/disrupting-AI-espionage>
- Anthropic. (2026, July 2). *More details on Fable 5's cyber safeguards and our jailbreak framework*. <https://www.anthropic.com/news/fable-safeguards-jailbreak-framework>
- CB Financial Services, Inc. (2026, May 11). *Form 8-K, Item 1.05: Material cybersecurity incidents* (Accession No. 0001605301-26-000021). U.S. Securities and Exchange Commission. <https://www.sec.gov/Archives/edgar/data/1605301/000160530126000021/cbfv-20260507.htm>
- Campbell, D., Kale, N., Sehwag, U. M., Herring, B., Knight, C. Q., Borges, D., Price, N., & Levinson, A. (2026). *Defensive refusal bias: How safety alignment fails cyber defenders* (arXiv:2603.01246). <https://arxiv.org/abs/2603.01246>
- Constantin, L. (2026, July 28). Hugging Face breach shows why incident response needs a multi-model AI strategy. *CSO Online*. <https://www.csoonline.com/article/4201361/>
- Darktrace. (2026). *State of AI cybersecurity 2026*. <https://www.darktrace.com/blog/state-of-ai-cybersecurity-2026-92-of-security-professionals-concerned-about-the-impact-of-ai-agents>
- European Commission. (2025, November 4). *AI Act: Commission publishes a reporting template for serious incidents involving general-purpose AI models with systemic risk*. <https://digital-strategy.ec.europa.eu/en/library/ai-act-commission-publishes-reporting-template-serious-incidents-involving-general-purpose-ai>
- European Union. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 55*. <https://artificialintelligenceact.eu/article/55/>
- Franceschi-Bicchierai, L. (2026, July 23). How AI guardrails are impeding the work of offensive cybersecurity researchers. *TechCrunch*. <https://techcrunch.com/2026/07/23/how-ai-guardrails-are-impeding-the-work-of-offensive-cybersecurity-researchers/>
- Hugging Face. (2026, July 16). *Security incident disclosure — July 2026*. <https://huggingface.co/blog/security-incident-july-2026>
- IBM & Ponemon Institute. (2026, July 29). *Cost of a data breach report 2026*. <https://newsroom.ibm.com/2026-07-29-ibm-study-one-in-four-malicious-breaches-are-ai-enabled,-costing-companies-6-million-on-average>

- Janakiram MSV. (2026, July 27). The Hugging Face breach exposed a gap in AI safety controls. *Forbes*. <https://www.forbes.com/sites/janakirammsv/2026/07/27/the-hugging-face-breach-exposed-a-gap-in-ai-safety-controls/>
- Jones, E. K., Dziemian, M., Fredrikson, M., & Kolter, J. Z. (2026). *A new framework for cybersecurity refusals in AI agents* (arXiv:2606.02644). <https://arxiv.org/abs/2606.02644>
- Larcher, H., Carreira, A., Rannou, C., et al. (2026, July 27). *Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident*. Hugging Face. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- Leclerc, J. (2026, July 1). *Octane joins Anthropic's Cyber Verification Program*. Octane Security. <https://www.octane.security/post/octane-joins-anthropics-cyber-verification-program>
- Legal IT Insider. (2026). *The five-day gap: What the OpenAI-Hugging Face incident should tell law firms*. <https://legaltechnology.com/the-five-day-gap-what-the-openai-hugging-face-incident-should-tell-law-firms/>
- Mandiant. (2026, March 23). *M-Trends 2026: Data, insights, and strategies from the frontlines*. Google Cloud. <https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026>
- Menlo Ventures. (2025, December 9). *2025: The state of generative AI in the enterprise*. <https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/>
- Nelson, A., Rekhi, S., Souppaya, M., & Scarfone, K. (2025, April). *Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile* (NIST SP 800-61r3). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-61r3>
- OpenAI. (2025, October 29). *Usage policies*. <https://openai.com/policies/usage-policies/>
- OpenAI. (2026a, April 14). *Trusted access for the next era of cyber defense*. <https://openai.com/index/scaling-trusted-access-for-cyber-defense/>
- OpenAI. (2026b, July 20). *Safety and alignment for long-horizon models*. <https://openai.com/index/safety-alignment-long-horizon-models/>
- OpenAI. (2026c, July 21). *Hugging Face model evaluation security incident*. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- Saltzer, J. H., & Schroeder, M. D. (1975). The protection of information in computer systems. *Proceedings of the IEEE*, 63(9), 1278-1308. <https://web.mit.edu/Saltzer/www/publications/protection/>
- U.S. Securities and Exchange Commission. (2023, July 26). *SEC adopts rules on cybersecurity risk management, strategy, governance, and incident disclosure by public companies* (Press release 2023-139). <https://www.sec.gov/news/press-release/2023-139>
- Wang, L., Chen, B., Ding, R., et al. (2026). *SecRespond: Benchmarking AI agents for real-world post-compromise incident response* (arXiv:2607.26791). <https://arxiv.org/abs/2607.26791>
- Wang, Z., Schiller, N., Li, H., Narayana, S. S., Nasr, M., Carlini, N., Qi, X., Wallace, E., Bursztein, E., Invernizzi, L., Thomas, K., Shoshitaishvili, Y., Guo, W., He, J., Holz, T., & Song, D. (2026). *ExploitGym: Can AI agents turn security vulnerabilities into real attacks?* (arXiv:2605.11086). <https://arxiv.org/abs/2605.11086>
- Willison, S. (2026, July 22). *OpenAI's accidental cyberattack against Hugging Face is science fiction that happened*. <https://simonwillison.net/2026/Jul/22/openai-cyberattack/>
- Wilson Sonsini Goodrich & Rosati. (2026). *"Shadow AI" triggers first SEC Form 8-K for unauthorized AI use: What financial institutions and public companies need to know*. <https://www.wsgr.com/en/insights/shadow-ai-triggers-first-sec-form-8-k-for-unauthorized-ai-use-what-financial-institutions-and-public-companies-need-to-know.html>
- Z.ai. (2026). *GLM-5.2 model card*. Hugging Face. <https://huggingface.co/zai-org/GLM-5.2>

Copyright, reuse, and citation

© 2026 Gerald C. Kane. This paper is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), <https://creativecommons.org/licenses/by-nc/4.0/>.

You may share and adapt it for noncommercial purposes, provided you credit Gerald C. Kane as the author, link to the license, and indicate any changes you made. You may not present this material, or adaptations of it, as your own work, and you may not sell it or use it in paid training or consulting without written permission.

Cite as: Kane, G. C. (2026). *Whose Permission Does Your Incident Response Need?*. <https://www.profkane.com/papers/incident-response-permission.pdf>

Permissions and commercial use: jerry@profkane.com